Ph.D. Title Proposal

on

Designing An Effective Data Storage Framework using Data Reduction Techniques

Submitted in partial fulfillment of the requirement
for the award of the Degree of

Doctor of Philosophy

in

Information Technology

by

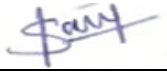
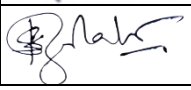
Mrs. Samita Atul Patil

under the guidance of

Dr. Savita Dinesh Sangam

Department of Information Technology
Mahavir Education Trust's
Shah and Anchor Kutchhi Engineering College
Chembur, Mumbai-400088
University of Mumbai
August 2024

PROPOSAL FOR TOPIC APPROVAL TO BE SUBMITTED TO THE UNIVERSITY OF MUMBAI FOR THE DEGREE OF **Ph.D. (SCIENCE AND TECHNOLOGY)** IN THE SUBJECT OF **INFORMATION TECHNOLOGY**

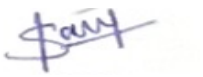
Title of Topic	:	Designing An Effective Data Storage Framework using Data Reduction Techniques
Name of the Student	:	Mrs. Samita Atul Patil
Basic Qualification	:	M.E (Computer Engineering) B.E.(Information Technology)
In the Case of Ph.D. student Year of Passing of PET/NET/SET/GATE (Teaching Experience)	:	PET - 2022
Name of the Research Supervisor/ Guide	:	Dr. Savita Dinesh Sangam
Date of Superannuation / Retirement	:	20/07/2033
Name of the Research Center	:	Shah and Anchor Kutchhi Engineering College
Signature of Student	:	
Signatue of the Research Supervisor	:	

Dr. Bhavesh Patel
Principal
Shah & Anchor Kutchhi Engg. College
Waman Tukaram Patil Marg,
Next to Dukes Co., Chembur,
Mumbai-400 088.

Sr. no.	Content	Page no.
1	Introduction	1
2	Motivation	2
3	Objectives	2
4	Literature Survey	3
5	Proposed Methodology	6
6	Expected Outcome	8
7	Conclusion	8
8	References	9

Statement by the Candidate

I wish to state that the work embodied in this synopsis titled “Designing An Effective Data Storage Framework using Data Reduction Techniques.” forms my own contribution to the work carried out under the guidance of Dr. Savita Dinesh Sangam at the Shah and Anchor Kutchhi Engineering College,Chembur Mumbai . I declare that this written submission represents my ideas in my own words and where others ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission.



(Candidate Signature)

Name: Mrs. Samita Atul Patil

Roll Number: 23DF18604IT001

Introduction

In today's data-driven world, the enormous amount of information generated each day is unexpected, generating big challenges to effective data storage and management. The rapid expansion of data across various domains ranging from social media and e-commerce to healthcare and scientific research necessitates innovative approaches to handle, store, and retrieve this data efficiently[1]. Traditional data storage methods, while foundational, are increasingly inadequate in managing the scale, velocity, and variety of modern data sets. This research delves into the realm of big data reduction methods as a pivotal solution for enhancing data storage efficiency and effectiveness[2] Big data reduction methods encompass a suite of techniques designed to minimize the data footprint without compromising data integrity and accessibility. Key among these techniques are data compression, data deduplication, and data preprocessing. Data compression algorithms reduce the size of data by eliminating redundancies and leveraging patterns within the data, thereby decreasing the storage space required[3]. Data deduplication further enhances storage efficiency by identifying and removing duplicate copies of repeating data, ensuring that only unique instances are stored [4]. Data cleaning, meanwhile, focuses on improving data quality by removing errors, inconsistencies, and irrelevant information, which not only optimizes storage but also enhances the accuracy and reliability of data analysis [5].A critical aspect of this research is addressing latency issues, particularly data ingestion latency and query latency. Data ingestion latency refers to the time taken to store incoming data, which can be a bottleneck in real-time data processing systems. By employing efficient data reduction methods, the system can streamline the ingestion process, thereby reducing latency and enabling faster data storage. Query latency, on the other hand, pertains to the time taken to retrieve and process data queries. Effective data reduction techniques can significantly speed up query response times by minimizing the amount of data that needs to be scanned and processed, leading to more efficient data retrieval and analysis[6] Another vital dimension of this research is the accuracy of data storage and retrieval. Ensuring data accuracy through comprehensive data cleaning, reduction, and deduplication processes is paramount. Accurate data storage not only enhances the reliability of analytical insights but also reduces the risk of errors in decision-making processes. By maintaining high data accuracy, organizations can trust the data they store and analyze, leading to better outcomes and more informed decisions[7].Storage efficiency is also a primary focus of this research, encompassing factors such as compression ratio, storage utilization, and throughput. A higher compression ratio indicates more efficient data storage, allowing organizations to store more data within the same physical space. Improved storage utilization ensures that the available storage resources are used optimally, minimizing wastage and maximizing capacity. Throughput, the rate at which data is processed and stored, is another critical metric that reflects the efficiency of data storage systems[8]. By enhancing throughput, organizations can handle larger volumes of data more effectively, supporting scalability and growth.This research leverages advanced algorithms and machine learning models to implement big data reduction methods, providing dynamic and adaptive solutions to varying data types and usage patterns. By integrating these technologies, the proposed system aims to achieve a balance between storage efficiency, data accuracy, and latency reduction. The ultimate goal is to revolutionize data storage practices, offering scalable, cost-effective, and robust solutions that meet the demands of contemporary data environments.In summary, this research on effective data storage using big data reduction methods addresses the pressing need for innovative solutions to manage the exponential growth of data. By focusing on reducing latency, ensuring accuracy, and optimizing storage efficiency, this study aims to provide a comprehensive framework for improving data storage practices[9]. Through the integration of advanced technologies and intelligent algorithms, the proposed methods offer a promising path toward more efficient, reliable, and scalable data storage solutions. The primary aim of this study is to explore and evaluate strategies

and techniques for reducing data storage requirements in big data environments. The study aims to identify effective methods to enhance storage efficiency, lower costs, improve system performance, and support sustainable data management practices. The purpose is to address challenges in managing and storing large volumes of data.The motivation behind achieving effective data storage using big data reduction methods lies in addressing critical performance and efficiency parameters. By focusing on latency, particularly data ingestion and query latency, the system can ensure faster data processing and retrieval, essential for real-time analytics and decision-making. Accuracy is another pivotal aspect, where enhanced data cleaning, dimentionality reduction, and deduplication techniques lead to higher quality data, minimizing errors and redundancies. Furthermore, optimizing storage efficiency through superior compression ratios and better storage utilization translates to significant cost savings and improved throughput[10]. These combined improvements not only bolster the overall performance and reliability of data systems but also provide a scalable and sustainable solution to managing vast amounts of data in diverse application domains.

Motivation

The rapid expansion of data in the digital age necessitates efficient storage frameworks that can handle vast volumes of information without compromising on speed and accessibility. Researching effective storage frameworks using various big data reduction methods is crucial to addressing this challenge. By exploring and developing advanced techniques such as data deduplication, compression algorithms, and dimentionality reduction methods, we can significantly reduce storage requirements and enhance data retrieval efficiency. This not only lowers costs associated with storage infrastructure but also improves data processing and analytics capabilities, enabling quicker decision-making and fostering innovation. Such research holds the potential to revolutionize industries reliant on big data, ensuring sustainable and scalable data management solutions for the future.

Objective

The primary aim of this study is to explore and evaluate strategies and techniques for reducing data storage requirements in big data environments. The study aims to identify effective methods to enhance storage efficiency, lower costs, improve system performance, and support sustainable data management practices. The purpose is to address challenges in managing and storing large volumes of data.

- To identify issues related to data storage systems.
- To propose an effective data storage architecture by applying data reduction technique.
- To find an optimal solution for data storage by evaluating the proposed system.
- To test the performance of the proposed system on real dataset.

Literature Review

The Hadoop Data Reduction Framework (HDRF) is a modular solution that improves data reduction techniques in Hadoop's Distributed File System (DFS). It offers minimal overhead and significant storage efficiency improvements across various datasets and workloads. Experimental results show HDRF achieves up to 48.96% better throughput and storage reduction compared to existing methods.[11]The paper highlights the importance of data quality in Big Data applications, particularly in weather monitoring and forecasting. It introduces a pre-processing framework to improve data quality, addressing issues like cleansing, noise filtering, integration, and transformation. The framework includes stages like consolidation, enhancement, transformation, reduction, discretization, and cleansing. It also incorporates data profiling and quality assessment to monitor and improve data quality throughout the lifecycle. The study emphasizes the need for continued research in data quality management to keep pace with Big Data challenges.[12]The paper discusses the challenges of securing large data storage in cloud environments, focusing on PPSecS, a solution that combines Salsa20, Lai-Massey cryptographic structure, and CP-ABE encryption. It presents a security analysis, shows efficient encryption and decryption compared to AES, and proposes future work to enhance data processing security and privacy. The authors also introduce a data reduction method for distributed and parallel processing.[13]The research paper presents SOPHON, a framework that reduces data traffic and training time in deep learning training by selectively offloading preprocessing tasks to remote storage servers. Implemented in Python on top of PyTorch, SOPHON can save 1.2-2.2x compared to existing solutions. It prioritizes images for maximum network traffic savings per unit of CPU time. SOPHON's potential lies in scenarios where remote I/O is a bottleneck.[14] The article discusses the challenges and improvements in smart grids (SGs) due to increased energy demand and technological advancements. SGs integrate data and communication technologies, enabling efficient resource usage. Big data analytics is crucial for managing energy demand and supply. The paper reviews data preprocessing methods, suggests machine learning and AI for SGs, and highlights challenges in preprocessing big data[15].The text discusses the use of bathymetric data from a multibeam echo sounder and hydrographic software. Two reduction methods, True Bathymetric Data Reduction (TBDRed) and Optimum Dataset (OptD), were developed to reduce large datasets while maintaining accuracy. The combination of these techniques improved visibility and efficiency, offering potential for hydrographic applications[16]The text discusses the use of bathymetric data from a multibeam echo sounder and hydrographic software. Two reduction methods, True Bathymetric Data Reduction (TBDRed) and Optimum Dataset (OptD), were developed to reduce large datasets while maintaining accuracy. The combination of these techniques improved visibility and efficiency, offering potential for hydrographic applications.[17]The article discusses the importance of preprocessing large TBM excavation statistics for machine learning. It introduces the drilling efficiency index and classifies inappropriate data into premature cycles, sensor defects, mechanical defects, human interruptions, and missing files. The report proposes a computer application, TBM-Processing, for recognizing, classifying, and modifying inappropriate data, ensuring a smooth database for machine learning[18].The research paper introduces a parallel preprocessing algorithm for analyzing large graphs using MapReduce to efficiently detect connected components. The algorithm reduces graph volume by over 99% in just two rounds, reducing execution time, network traffic, and the number of MapReduce rounds needed for connected component detection. This method improves network efficiency and allows for faster analysis of large datasets, making it an efficient method for connected component detection[19].The document explores data reduction techniques in additive manufacturing analysis, evaluating Blosc, SZ, ZFP, and Hybrid Data Sampling. It identifies Blosc for data-free applications, SZ and ZFP for accuracy, and Hybrid Data Sampling for speed-critical scenarios. SZ has the highest reduction ratio but trades compres-

sion ratio and data quality. The choice depends on application requirements. Reduction ratio () is used to quantify the amount of data reduced, with a ratio of 2 indicating a 50% reduction[20].The report introduces a Low-Overhead Compressibility Prediction method for high-performance lossless data compression, aiming to reduce computational load and address storage space and I/O bandwidth issues. It evaluates factors affecting compression ratio and discusses hardware implementation issues [21]. Wolfgang Erb introduces graph wedgelets, a method for compressing graph records using adaptive binary wedge splits. The technique uses recursive wedge splits to obtain sparse graph signals. Erb introduces learning algorithms for binary wedge partitioning trees and presents examples of their effectiveness in compressing signals and images [22]. The document introduces a Blockchain-based Special Keys Security Model (BSKM) with a Path Compression Algorithm for Big Data, addressing privacy challenges in data collection. The model ensures confidentiality, integrity, and consistency, employing a special key structure and a revocation mechanism. The experimental evaluation compares BSKM with traditional security models, highlighting its effectiveness in data reduction and access[23]. The paper explores data compression strategies, their evolution from Morse code to modern techniques, and their applications in text, image, audio, and biomedical signal compression. It categorizes techniques based on data quality, coding schemes, data type, and applications, and evaluates their performance using metrics like Compression Ratio and PSNR [24]. The text discusses managing large data volumes and the importance of data deduplication. It reviews various techniques, focusing on the optimized content-defined chunking method. The paper also discusses a scalable bucket indexing method for reducing hash judgment time on the Hadoop Distributed File System. The authors acknowledge support from colleagues and academic advisors[25]. The document introduces a new method for inline data deduplication in storage systems, InDe, which addresses data fragmentation and improves overall performance. It uses a learning algorithm to adjust old container references, VCRC metric, and two rewrite algorithms, F-greedy and F-grasping+. This advancement in data deduplication addresses challenges and limitations [26].The report highlights the importance of efficient digital data storage in the virtual era, emphasizing data deduplication as a cost-effective and efficient technique. It explores various deduplication techniques, including source, target, inline, post-process, content-based, and chunk-level deduplication, and discusses the challenges of storing large volumes of data. The report provides practical insights for cloud storage providers and organizations. It highlights the advantages of deduplication in terms of cost-effectiveness, storage space optimization, and bandwidth utilization, with the potential to control up to 60% of storage space wastage[27]. The report examines deduplication algorithms in supply models with probabilistic substitutions, focusing on fixed-length deduplication, modified and adaptive deduplication, and variable-length deduplication. It reveals that the performance of the mFLD algorithm depends on bite length, and discusses the use of various techniques, their computational performance, and their suitability for specific source models[28].The document explores records deduplication techniques in cloud computing for efficient storage management of increasing virtual data. It discusses challenges of data duplication and potential benefits, categorizing techniques and discussing their pros and cons. It also covers the architecture of a data deduplication system, including client-server, metadata-server, and data-server components.[29]. The study examines the impact of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) on machine learning algorithms' performance using three datasets. The authors suggest a methodology that includes feature engineering for improved data quality and compares ML algorithm performance with and without these techniques, indicating their potential for large data analysis[30].The text explores the use of PCA and LDA in enhancing machine learning algorithm performance. It compares the performance of algorithms with and without dimensionality reduction, finding PCA more effective than LDA. The study also suggests that Decision Tree and Random Forest classifiers aren't significantly affected. It emphasizes feature engineering in pre-processing and their potential in large data

analysis[31]. The text introduces a Word Pattern Clustering (WPC) approach that uses a Similarity Function for clustering and a Transformation Matrix for dimensionality reduction. The WPC method is evaluated using metrics like Information-Ratio for Dimension Reduction, Accuracy, and Recall, and compared to existing processes like LDA-SVM and AFO-MKSVM. The WP-SC method offers benefits in DR, classification accuracy, and scalability, making it a valuable tool for big data analysis[32].The ”curse of dimensionality” is a challenge in handling high-dimensional data, causing increased storage and computational costs. Feature dimensionality reduction aims to map data from high-dimensional spaces to lower ones while preserving important characteristics. Techniques include feature selection and extraction, using algorithms like filter, wrapper, and beam search[33].A new algorithm for reducing computational complexity in unsupervised learning for dimensionality reduction of smart meter data uses K-means clustering, a new distance metric, and a stochastic-based approach, evaluated using real consumption data[34].The Non-Linear Embeddings Surveyor (NoLiES) is a solution for explaining non-linear embeddings of high-dimensional data using rangesets. It improves interpretability through case studies and real-world applications in thermodynamics. NoLiES enhances interpretability across various fields and references dimensionality reduction methods for easier visualization and analysis[35]. The national status of optimizing storage using big data reduction methods by machine learning is active, with enterprises and research institutions adopting advanced techniques for efficient data management and processing. This enables intelligent and scalable data solutions in various fields.Machine learning-based methods like deep learning, distributed and parallel learning, transfer learning, and active learning are being increasingly adopted in industries like finance, healthcare, and telecommunications to optimize storage, reduce costs, and enhance data analysis capabilities, despite the complexities of big data. Organizations face challenges in managing big data due to high costs, slow retrieval times, and scalability issues. Traditional storage systems lack adaptability and efficiency. Big data reduction methods and machine learning algorithms can improve system performance. Hybrid storage solutions combine on-premises and cloud storage for flexibility. Comprehensive research is needed to evaluate these solutions’ practical implications, performance, cost savings, and scalability benefits.

Proposed Methodology

To create an efficient data storage and retrieval system that can handle both structured and unstructured data, its crucial to implement a comprehensive workflow that encompasses data ingestion, preprocessing, reduction, deduplication, compression, metadata management, and dynamic retrieval. This approach ensures optimal storage utilization while maintaining data integrity and accessibility.

1. Data Ingestion Layer The purpose of data ingestion is to collect and import data from various sources into a storage system. This process involves identifying sources, using APIs or data streaming platforms, and ensuring data integration into a unified system, which may involve real-time or batch processing. Ingestion of structured and unstructured data involves reading and validating data, handling missing values, normalizing data, and transforming it into a suitable format for storage, and preprocessing binary files like images and videos to ensure they meet required format and quality standards[36-37].

2. Data Reduction Layer The data reduction function reduces data volume while maintaining integrity and information content, making it easier to store and process. Techniques like Principal Component Analysis or t-SNE reduce feature count while retaining information.Data reduction techniques for structured and unstructured data involve dimensionality reduction, resizing images or videos, and reducing resolution, frame rate, or video format to minimize data size while retaining essential quality attributes, such as PCA for structured data and resizing for unstructured data.[38-40]

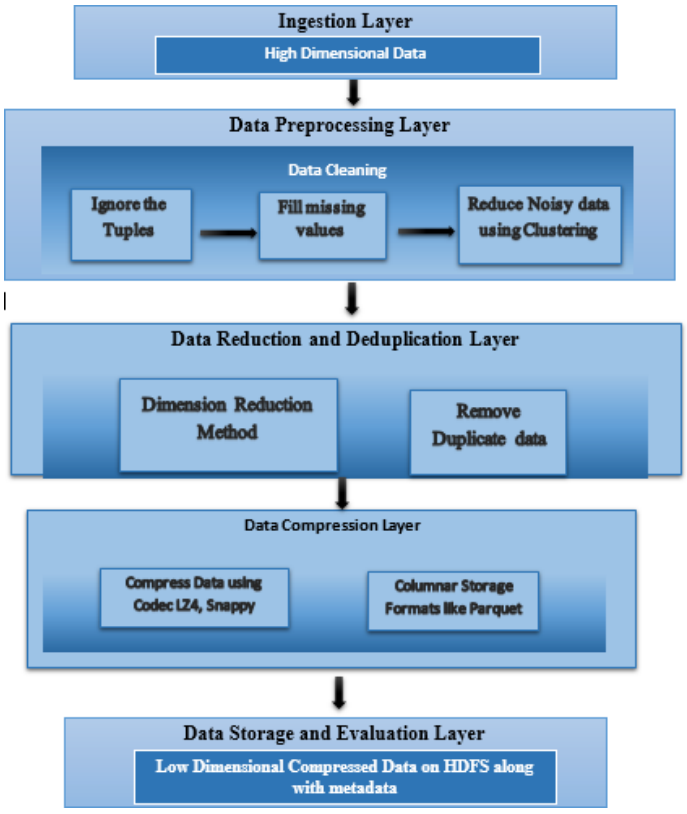

Ingestion Layer
High Dimensional Data
Data Preprocessing Layer
Data Cleaning
Ignore the Tuples
Fill missing values
Reduce Noisy data using Clustering
Data Reduction and Deduplication Layer
Dimension Reduction Method
Remove Duplicate data
Data Compression Layer
Compress Data using Codec LZ4, Snappy
Columnar Storage Formats like Parquet
Data Storage and Evaluation Layer
Low Dimensional Compressed Data on HDFS along with metadata

Figure 1: system flow diagram .

3. Data Deduplication Deduplication is the process of identifying and removing duplicate copies of data. This step is crucial for optimizing storage space, ensuring that only unique instances of data are stored. Deduplication algorithms compare new data with existing data to detect duplicates and maintain only one copy[41-42].

4. Data Compression Layer The Data Compression Layer is a layer that efficiently encodes data, saving storage space and speeding up data transfer. It consists of Lossless Compression algorithms like Huffman Coding and Lossy Compression techniques, which allow some loss of data fidelity in exchange for higher compression ratios, often used in scenarios where perfect reconstruction isn't necessary. Compression techniques like gzip and Snappy reduce file size and allow exact reconstruction during retrieval for structured data. Efficient, lossless compression algorithms like PNG and lossless codecs minimize storage footprint for unstructured data, balancing quality loss with data integrity[43-44].

5. Metadata Management and Storage Metadata management involves storing essential information about the original data attributes and any transformations applied during preprocessing, reduction, and compression. This includes details like original dimensions, formats, compression ratios, and any applied reductions. Metadata ensures that the data can be accurately reconstructed during retrieval.The reduced, deduplicated, and compressed data, along with its metadata, is stored in a distributed file system such as Hadoop Distributed File System (HDFS). HDFS provides scalable and reliable storage that can handle large volumes of data efficiently[45-46].

6. Output Layer Retrieval for structured and unstructured data involves accessing compressed and deduplicated data, using metadata to decompress and reconstruct the original data structure. This includes restoring original dimensions and removed features. For unstructured data, decompression is used and reductions reversed, ensuring accurate restoration of quality and format using metadata.[47-48]
The output layer is responsible for delivering processed data to end-users or downstream applications in a usable format. It comprises components such as data retrieval, visualization and reporting, and API and interfaces. Data retrieval efficiently queries and retrieves data using a compressed structure.[49] Visualization and reporting transform data into charts or graphs, making it easier for users to interpret and act upon[50].

Expected Outcomes

- Identification of the most effective data reduction methods for optimized storage.
- Implementation of an effective big data storage architecture by applying various data reduction techniques.
- Demonstration of the benefits of combining multiple data reduction techniques for improved data storage.
- Performance testing of the proposed system on real dataset.

Conclusion

In conclusion, the challenge of managing and storing the ever-growing volume of data in today's digital age necessitates innovative solutions to optimize storage efficiency and performance. The existing systems often struggle with scalability and cost-effectiveness, leading to inefficiencies in data handling and retrieval. The proposed system leverages advanced big data reduction methods to enhance data storage by compressing and deduplicating data, thereby significantly reducing the storage footprint and improving access times. Unlike previous research, this system integrates cutting-edge algorithms and machine learning techniques to dynamically adjust to varying data types and usage patterns, ensuring optimal performance and adaptability. This approach not only mitigates the limitations of traditional storage systems but also provides a robust, scalable, and cost-effective solution for effective data management in big data environments. Compared to other approaches, our system dynamically adapts to various data types and usage patterns through the use of sophisticated algorithms and machine learning models, providing superior performance and flexibility. Future research can further enhance these methods by exploring real-time data processing capabilities, integrating more advanced machine learning techniques for predictive analytics, and developing more robust frameworks for handling even larger and more complex data sets, ultimately paving the way for more efficient and effective data storage solutions in an increasingly data-driven world.

References

[1] Theodorakopoulos, L.; Theodoropoulou, A.; Stamatiou, Y. A State-of-the-Art Review in Big Data Management Engineering: Real-Life Case Studies, Challenges, and Future Research Directions Eng 2024, 5, 1266-1297.

[2] Khoei, T.T., Singh, A. Data reduction in big data: a survey of methods, challenges and future directions Int J Data Sci Anal(2024).

[3] Xiaoyun Li, Hongyu Zhang, Van-Hoang Le, and Pengfei Chen. 2024. LogShrink: Effective Log Compression by Leveraging Commonality and Variability of Log Data. In Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (ICSE '24). Association for Computing Machinery, New York, NY, USA, Article 23, 112.

[4] K. Vijayalakshmi and V. Jayalakshmi, "Analysis on data deduplication techniques of storage of big data in cloud," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 976-983.

[5] F. Zou, "Research on data cleaning in big data environment," 2022 International Conference on Cloud Computing, Big Data and Internet of Things (3CBIT), Wuhan, China, 2022, pp. 145-148.

[6] Y. Wang, Y. Li, J. Sui and Y. Gao, "Data Factory: An Efficient Data Analysis Solution in the Era of Big Data," 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), Xiamen, China, 2020, pp. 28-32.

[7] J. Guo and Y. Cui, "Research on Big Data Retrieval System of Intelligent Computer AI in Post-Moore Era," 2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 2024, pp. 1231-1235.

[8] W. Wan et al., "Compressed Data Direct Computing for Databases," in IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 5, pp. 1902-1918, May 2024.

[9] Y. Kim et al., "Low-Overhead Compressibility Prediction for High-Performance Lossless Data Compression," in IEEE Access, vol. 8, pp. 37105-37123, 2020.

[10] 10.C. Jaranilla, H. Shin, S. Yoo, S. -j. Cho and J. Choi, "Tiered Storage in Modern Key-Value Stores: Performance, Storage-Efficiency, and Cost-Efficiency Considerations,"2024 IEEE International Conference on Big Data and Smart Computing (BigComp), Bangkok, Thailand, 2024, pp. 151-158.

[11] R. N. S. Widodo, H. Abe and K. Kato, "Hadoop Data Reduction Framework: Applying Data Reduction at the DFS Layer," in IEEE Access, vol. 9, pp. 152704-152717, 2021.

[12] A. Juneja and N. N. Das, "Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application," 2020 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2020, pp. 559-563.

[13] Bouleghlimat, I., Boudouda, S. Hacini, S. PPSecS: Privacy-Preserving Secure Big Data Storage in a Cloud Environment. Arab J Sci Eng 49, 32253239 (2024).

[14] Meng Wang, Gus Waldspurger, and Swaminathan Sundararaman. 2024. A Selective Preprocessing Offloading Framework for Reducing Data Traffic in DL Training. In Proceedings of the 16th ACM Workshop on Hot Topics in Storage and File Systems (HotStorage '24). Association for Computing Machinery, New York, NY, USA, 6370.

[15] T. A. Alghamdi and N. Javaid, "A Survey of Preprocessing Methods Used for Analysis of Big Data Originated From Smart Grids," in IEEE Access, vol. 10, pp. 29149-29171, 2022

[16] Wlodarczyk-Sielicka, M.; Blaszczak-Bak, W. Processing of Bathymetric Data: The Fusion of New Reduction Methods for Spatial Big Data.Sensors-2020.

[17] V. etin and O. Yldz, A comprehensive review on data preprocessing techniques in data analysis,Pamukkale niversitesi Mhendislik Bilimleri Dergisi, vol. 28, no. 2, pp. 299312, 2022.

[18] Hao-Han Xiao; Wen-Kun Yang; Jing Hu, Significance and methodology: Preprocessing the big data for machine learning on TBM performance, ISSN: 2467-9674, Vol: 7, Issue: 4, Page: 680-701,2022.

[19] R. Abdolazimi, M. Heidari, A. Esmaeilzadeh and H. Naderi, "MapReduce Preprocess of Big Graphs for Rapid Connected Components Detection," 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2022, pp. 0112-0118, doi: 10.1109/CCWC54503.2022.9720798..

[20] C. Nichols, M. H. Fulp, N. DeBardeleben and J. C. Calhoun, "Exploring Data Reduction Techniques for Additive Manufacturing Analysis," 2022 IEEE/ACM 8th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD), Dallas, TX, USA, 2022, pp. 21-28

[21] Y. Kim et al., "Low-Overhead Compressibility Prediction for High-Performance Lossless Data Compression," in IEEE Access, vol. 8, pp. 37105-37123, 2020.

[22] W. Erb, "Graph Wedgelets: Adaptive Data Compression on Graphs Based on Binary Wedge Partitioning Trees and Geometric Wavelets," in IEEE Transactions on Signal and Information Processing over Networks, vol. 9, pp. 24-34, 2023.

[23] C. Bakir, "New Blockchain Based Special Keys Security Model With Path Compression Algorithm for Big Data," in IEEE Access, vol. 10, pp. 94738-94753, 2022.

[24] Uthayakumar, J., Vengattaraman, T., Dhavachelvan, P. (2018). A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications. Journal of King Saud University - Computer and Information Sciences.

[25] Kumar, N., Shobha, Jain, S.C. (2020). Efficient Data Deduplication for Big Data Storage Systems. In: Panigrahi, C., Pujari, A., Misra, S., Pati, B., Li, KC. (eds) Progress in Advanced Computing and Intelligent Engineering. Advances in Intelligent Systems and Computing, vol 714. Springer, Singapore

[26] Lifang Lin, Yuhui Deng, Yi Zhou, and Yifeng Zhu. 2023. InDe: An Inline Data Deduplication Approach via Adaptive Detection of Valid Container Utilization. ACM Trans. Storage 19, 1, Article 6 (February 2023), 27 pages.

[27] G. Sujatha, Jeberson Retna Raj, A Comprehensive Study of Different Types of Deduplication Technique in Various Dimensions, International Journal of Advanced Computer Science and Applications(IJACSA), Volume 13 Issue 3, 2022.

[28] H. Lou and F. Farnoud, ”Data Deduplication With Random Substitutions,” in IEEE Transactions on Information Theory, vol. 68, no. 10, pp. 6941-6963, Oct. 2022.

[29] K. Vijayalakshmi and V. Jayalakshmi, ”Analysis on data deduplication techniques of storage of big data in cloud,” 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 976-983.

[30] Vogelstein, J.T., Bridgeford, E.W., Tang, M. et al. Supervised dimensionality reduction for big dataNat Commun 12, 2872 (2021).

[31] G. T. Reddy et al., ”Analysis of Dimensionality Reduction Techniques on Big Data,” in IEEE Access, vol. 8, pp. 54776-54788, 2020.

[32] Jijo Varghese, P. Tamil Selvan. (2023). A Novel Clustering and Matrix Based Computation for Big Data Dimensionality Reduction and Classification. Journal of Advanced Research in Applied Sciences and Engineering Technology, 32(1), 238251.

[33] Jia, W., Sun, M., Lian, J. et al. Feature dimensionality reduction: a review. Complex Intell. Syst. 8, 26632693 (2022).

[34] M. Mohajeri, A. Ghassemi and T. A. Gulliver, ”Fast Big Data Analytics for Smart Meter Data,” in IEEE Open Journal of the Communications Society, vol. 1, pp. 1864-1871, 2020.

[35] J.-T. Sohns, M. Schmitt, F. Jirasek, H. Hasse and H. Leitte, ”Attribute-based Explanation of Non-Linear Embeddings of High-Dimensional Data,” in IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 1, pp. 540-550, Jan. 2022.

[36] Khan, Samiya and Mansaf Alam. Preprocessing framework for scholarly big data management. Multimedia Tools and Applications 82 (2022): 39719 - 39743.

[37] Albattah, W.; Khan, R.U.; Khan, K. ”Attributes Reduction in Big Data” Appl. Sci. 2020, 10, 490.

[38] Fernandes, V.; Carvalho, G.; Pereira, V.; Bernardino, J. Analyzing Data Reduction Techniques: An Experimental Perspective. Appl. Sci. 2024, 14, 3436.

[39] Haeri, Hossein et al. Adaptive Granulation: Data Reduction at the Database Level. International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (2023).

[40] Gyamerah, Seth et al. Comparative Analysis of Feature Extraction of High Dimensional Data Reduction Using Machine Learning Techniques. American Journal of Electrical and Computer Engineering (2023)

[41] Yang, F.; Liu, S.; Dobriban, E.;Woodruff, D.P. How to Reduce Dimension with PCA and Random Projections? IEEE Trans. Inf. Theory 2021, 67, 81548189.

[42] ARORA, R., VETRITHANGAM, D. (2024). ADVANCEMENTS IN DEDUPLICATION TECHNIQUES FOR EFFICIENT DATA STORAGE. Journal of Theoretical and Applied Information Technology, 102(5).

[43] Jeong, He Bum et al. Lossless Data Compression with Bit-back Coding on Massive Smart Meter Data. 2022 IEEE International Conference on Big Data (Big Data) (2022): 6667-6669.

[44] Harb, H.; Jaoude, C.A. Combining compression and clustering techniques to handle big data collected in sensor networks. In Proceedings of the 2018 IEEE Middle East and North Africa Communications Conference, MENACOMM, Jounieh, Lebanon, 1820 April 2018; pp. 16.

[45] Li, Chris, and Jane Smith. ”Master data management in the era of big data: challenges and opportunities.” (2024).

[46] Safiatou S. Sore, T. Frederic T.Frdric Ouedraogo, Moustapha M. Bikienga, and Yaya Y. Traore. 2024. Towards a More Generic and Elastic Metadata Management Model in a Data Lake Environment. In Proceedings of the 2024 16th International Conference on Machine Learning and Computing (ICMLC ’24). Association for Computing Machinery, New York, NY, USA, 4451.

[47] X. Huang, W. Yi, J. Wang, Z. Xu, Hadoop-based medical image storage and access method for examination series, Math. Probl Eng. (3) (2021) 110, 2021.

[48] Hariharan, R., R. Vedapradha, and Sudha E. Megha Pandey. ”Pilgrim Tourism: Tourist Experience in Accommodation Using Content Analysis.” Journal of Informatics Education and Research 4.2 (2024).

[49] Keizer, Navin, Onur Ascigil, Michal Krl, Dirk Kutscher, and George Pavlou. ”A Survey on Content Retrieval on the Decentralised Web.” ACM Computing Surveys 56, no. 8 (2024): 1-39.

[50] S. Wang, S. Zhong, Y. Hong and Y. Zheng, ”Big Data Analysis Method for Power Information Based on Visualization Technology,” 2022 2nd International Conference on Networking Systems of AI (INSAI), Shanghai, China, 2022, pp. 255-259, doi: 10.1109/INSAI56792.2022.00055.